Accompanying technology development in the Human Brain Project: From foresight to ethics management

Authors

Dr Christine Aicardi, Department of Global Health & Social Medicine, King's College London, United Kingdom, christine.aicardi@kcl.ac.uk (corresponding author)

Dr B. Tyr Fothergill, Centre for Computing and Social Responsibility, De Montfort University, United Kingdom, tyr.fothergill@dmu.ac.uk

Dr Stephen Rainey, Uehiro Centre for Practical Ethics, University of Oxford, United Kingdom, stephen.rainey@philosophy.ox.ac.uk

Prof. Bernd Carsten Stahl, Centre for Computing and Social Responsibility, De Montfort University, United Kingdom, bstahl@dmu.ac.uk

Dr Emma Harris, Centre for Computing and Social Responsibility, De Montfort University, United Kingdom, emma.harris@dmu.ac.uk

Abstract

This paper addresses the question of managing the existential risk potential of general Artificial Intelligence (AI), as well as the more near-term yet hazardous and disruptive implications of specialised AI, from the perspective of a particular research project that could make a significant contribution to the development of Artificial Intelligence (AI): the Human Brain Project (HBP), a ten-year Future and Emerging Technologies Flagship of the European Commission. The HBP aims to create a digital research infrastructure for brain science, cognitive neuroscience, and brain-inspired computing. This paper builds on work undertaken in the HBP's Ethics and Society subproject (SP12). Collaborators from two activities in SP12, Foresight and Researcher Awareness on the one hand, and Ethics Management on the other, use the case of machine intelligence to illustrate key aspects of the dynamic processes through which questions of ethics and society, including existential risks, are

approached in the organisational context of the HBP. The overall aim of the paper is to provide practice-based evidence, enriched by self-reflexive assessment of the approach used and its limitations, for guiding policy makers and communities who are, and will be, engaging with such questions.

Keywords: Responsible Research and Innovation; Human Brain Project; foresight; ethics management; Artificial Intelligence

Acknowledgements

We thank our colleagues and collaborators in the Human Brain Project, most especially the Ethics and Society subproject (SP12), and the anonymous reviewers for their valuable feedback. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720270 (HBP SGA1). The ideas presented represent the views of the authors and we do not claim to reflect the position of the Human Brain Project or of its funder, the European Commission.

1 Introduction

Existential risks, 'x-risks' for short, are commonly understood as hypothetical future events that could cause the extinction of humanity or drastically alter its continued existence. The existential risks associated with technological developments have attracted much attention in the recent past, with the creation of dedicated institutions such as the Future of Life Institute (founded in 2014),[1] or the Centre for the Study of Existential Risk at the University of Cambridge (founded in 2012),[2] a concern of which is to bring together "the 'x-risk ecosystem' – a thriving community of researchers and others, inside and outside academia, united by a common interest in potential serious hazards of powerful and beneficial new technologies […] to ask ourselves where our efforts should best be directed, over the rest of the decade and beyond."[3]

This paper aims to contribute to the debate by focusing on the risks (not all existential yet no less serious) posed by one such technology, artificial intelligence (AI). The topic of machine intelligence as a potential threat to humanity is not new. It has long been a theme in popular culture, the archetypal mad scientists Faust and Frankenstein established the powerful trope of pessimism about scientific endeavours and a fear of their results (Weingart, 2010, p. 339). Modern cinema has tended to reinforce these concerns, particularly when depicting machine intelligence. Films such as *Terminator* (dir. Cameron, 1984), *The Matrix* (dir. Wachowskis, 1999) and *Transcendence* (dir. Pfister, 2014) depict machine intelligence as dangerous and destructive, though it should be noted that several recent films and TV series have been more ambiguous in this regard and that video games such as *Mass Effect: Andromeda* (dev. BioWare, 2017) are optimistic in their depictions.

These cultural trends may help to explain why the Special Eurobarometer 382: Public Attitudes towards Robots (2012)[4] found such negative attitudes to AI and robotics in 'human' roles. The survey found that a large majority of respondents were sceptical or fearful of machine intelligence becoming part of their personal, as opposed to professional, lives. '[T]here is widespread agreement that robots should be banned in the care of children, the elderly or the disabled (60%) with large minorities also wanting a ban when it comes to other 'human' areas such as education (34%), healthcare (27%) and leisure (20%)' (2012: 4).

[1] https://futureoflife.org/, consulted 10/01/2018.
[2] https://www.cser.ac.uk/, consulted 10/01/2018.
[3] http://www.crassh.cam.ac.uk/events/27021, consulted 19/12/2017.
[4] Special Eurobarometer 382: Public Attitudes towards Robots, Conducted by TNS Opinion & Social at the request of Directorate-General for Information Society and Media (INSFO), Survey co-ordinated by Directorate-General Communication, 2012.

More recently, the topic has attracted a high level of attention, as indicated by the UK Parliament's Science and Technology Select Committee's report on "robotics and artificial intelligence" (House of Commons Science and Technology Committee, 2016) which mirrors reports on the same topic from the US (Executive Office of the President, 2016a, 2016b) and the European Parliament (Committee on Legal Affairs, 2017). This heightened attention by policymakers reflects a growing awareness that the confluence of artificial intelligence techniques, big data, high processing power at low energy cost, and the increasing spread of information and communication technologies (ICTs) has arrived at the point where it can plausibly be said to have potentially significant impact on people's lives.

This growing awareness of the increasing power of AI does not by itself imply that these technologies pose a particular risk, even less that they pose an existential risk in the sense that they threaten the very survival of humanity or at least of our current way of life. They are nevertheless a good starting point to ask whether such risks may materialise and how they could be addressed.

This paper addresses that question from the perspective of a research project with the potential to make a significant contribution to the development of AI. The Human Brain Project (HBP, www.humanbrainproject.eu), a ten-year Future and Emerging Technologies Flagship initiative of the European Commission, has the overall aim to create an ICT-based scientific research infrastructure for brain research, cognitive neuroscience, and brain-inspired computing. To this end, it brings together a number of activities, including animal, human, cognitive, and theoretical neuroscience as well as platform development in the fields of neuroinformatics, high performance analytics and computing, medical informatics, neuromorphic computing, and neurorobotics. This combination of activities offers the possibility of ground-breaking insights that can substantially change or accelerate the development of artificial intelligence. The exact capabilities of these new technologies are still difficult to assess, but in seeking to capitalise on our understanding of animal and human brains, we have high expectations regarding their impact. The flipside of these hopes for the development of novel technologies is that they may constitute risks that are difficult or even impossible to evaluate.

From early on in its development, the HBP has been aware of these and other social, ethical and philosophical concerns, and has dedicated a set of activities to such questions. These are organised around the principles of Responsible Research and Innovation (RRI). RRI, in the interpretation adopted by the UK Engineering and Physical Research Council (EPSRC) through its AREA framework (Anticipate, Reflect, Engage, Act),[5] suggests that research needs to include anticipation of possible future consequences, reflection on the rationale and justification of research, engagement with

[5] https://www.epsrc.ac.uk/research/framework/area/, consulted 19/12/2017.

various stakeholders, and translation of these activities into action. It is hoped that incorporating these principles in all aspects of the research and innovation process will make it more socially responsible. The HBP has implemented these principles through four work packages in Subproject SP12, Ethics and Society, which cover foresight and researcher awareness, conceptual and philosophical reflection, public engagement, and ethics management. Although existential risks are unprecedented and thus particularly difficult to identify, these activities will hopefully detect if the HBP starts raising existential risks, and in any case recognize other serious risks, before recommending suitable ways of addressing them, and developing appropriate action plans.[6]

This present paper builds on work undertaken in the HBP's Ethics and Society subproject (SP12). Collaborators from two tasks and a work package in SP12, Foresight, Researcher Awareness, and Ethics Management use the case of machine intelligence to illustrate key aspects of the dynamic process through which questions of ethics and society, including existential risks, are approached in the HBP organisation. The overall aim of the paper is to provide practice-based evidence, enriched by the self-reflexive assessment of the approach used and its limitations, for guiding policy makers and communities who are, and will be, engaging with such questions.

The foundational work was initiated in the Ramp-Up Phase of the HBP (between October 2013 and March 2016) and continues into the 1st tranche of the HBP Operational Phase (SGA1, between April 2016 and March 2018), around the potential contribution that the Project could make to future computing and robotics, machine intelligence in particular.

Firstly, this paper briefly presents the conclusions of the foresight work that was conducted during the Ramp-Up Phase of the HBP. It then details how the resulting recommendations are being developed for action by the Ethics Management and the Researcher Awareness teams. It thereby demonstrates how researcher awareness and ethics management can evaluate and act on the issues initially raised in foresight work, and take them back to the researchers and other members of the HBP in order to increase their capacity to reflect on ethical, social, and regulatory issues, thus helping close the loop between anticipation and action in the AREA framework. Finally, the paper reflects on the overall process from a methodological perspective, its benefits and limitations.

This paper makes an important contribution to knowledge by demonstrating how principles of RRI can be integrated into large multidisciplinary and international projects, allowing for a continuous risk assessment. Furthermore, our work confirms that AI has the potential to have socially disruptive consequences that are in need of attention. It contributes to the debate in RRI by highlighting the

[6] For detailed perspectives on the role and activities of the Ethics and Society Subproject in the HBP, see (Aicardi et al., 2017), (Evers, 2017), (Rainey et al., 2017).

importance of practical management-oriented activities in fulfilling the aspirations of RRI. In practical terms, it provides an example of structures and processes that can be employed to integrate continuous self-reflection into research projects. This will be a useful resource which other AI projects, and projects with potential existential risks more generally, can learn from and develop further.

2 HBP potential for developments in machine intelligence: Foresight analysis

The Foresight Lab[7] of the Human Brain Project is a research group comprised of social scientists, whose aim is to evaluate potential social, ethical, legal, and economic consequences of new knowledge and technologies produced from the work of the HBP, thus addressing the AREA framework dimension of 'anticipation'. In order to do this, members conduct systematic foresight exercises on key aspects of the HBP to identify and evaluate these potential impacts. During the Ramp-Up Phase of the HBP, the Foresight Lab conducted three such exercises, one for each of the main areas to which the HBP is expected to make significant contributions: neuroscience, medicine, and computing and robotics. The foresight work done in relation to the latter area is of particular relevance to the present paper.

2.1 Methodology

During the Ramp-Up phase of the HBP, the Foresight Lab analysed the material collected through various activities in SP12 to produce a Foresight Report on Future Computing and Robotics, in relation to the developments that could be expected from the HBP. What follows draws directly from this report, to which we refer the reader for further details.[8]

As for its other foresight exercises, the overall methodological strategy used by the Foresight Lab was to collect and analyse the views and perspectives of key stakeholders with qualitative research tools from the empirical social sciences. To that end, an extensive period of 'horizon scanning' was initiated; this involved examining the literature, both academic and popular, and identifying key themes and questions. Then, a number of important themes were distilled from the 'horizon scanning' work and complemented by interview and survey data collected by the researcher awareness team as part of their work on ethical and social perceptions in the HBP. These themes were then used as the basis for two webinars with key stakeholders. These webinars were co-

[7] The HBP Foresight Lab, headed by Prof. Nikolas Rose, belongs to the Department of Global Health & Social Medicine at King's College London, UK (http://www.kcl.ac.uk/sspp/departments/sshm/research/Research-Groups/BIOS/BIOS-Projects/HBP/The-Human-Brain-Project.aspx, consulted 19/12/2017).

[8] For the full online version of the report, see (Rose et al., 2016).

organized with the Danish Board of Technology Foundation, who coordinate the main citizen and stakeholder engagement activities in SP12. The webinars, open to an invited audience of 25-35 persons, focussed on the themes of dual use (military/civilian), intelligent machines, human-robot interaction, machine-learning, and brain computer interfaces. Participants were invited to ask questions about future directions, potential alternative pathways, and the risks and benefits of each aspect. The webinar presentations and ensuing debates were recorded for subsequent analysis. Furthermore, participants were encouraged to continue the debate after the webinars, and an extensive, effective email discussion on the themes of machine intelligence and machine consciousness took place between key researchers in the HBP and external experts, which, thanks to their consent, provided additional material for analysis.

Regarding the methodological approach adopted for analysing the data collected, a major challenge was that the HBP could contribute significantly to many areas of ICT and robotics, leading to a wide variety of future applications which cross-cut many domains. Confronted with such an extensive range of possible products and domains of application, the Foresight Lab decided against taking an inventory-like approach to each kind of product and domain, because this would generate a long and disparate catalogue, hindered by its ability to identify salient common features. Indeed, this would have been counter-productive in the context of foresight exercises, which aim to anticipate future trends and drivers. Instead, the authors of the report adopted a holistic approach, looking at hardware and software, machines and humans, as parts of larger systems, with the aim of identifying the key social and ethical challenges posed by the potential contributions of the HBP to future ICT and robotics. Two cross-cutting, and related, key themes eventually stood out: human-machine integration, which falls somewhat outside the scope of the present discussion; and intelligent machines, the primary focus of this paper.

As a last methodological clarification, it is important to indicate that the view taken in the report is that the growing amount of speculation about general Artificial Intelligence (with some suggesting that we will soon reach the 'singularity', the point at which machine intelligence overtakes human intelligence) is somewhat premature, and diverts our attention from more pressing social and ethical issues arising in connection to the proliferation and rapidly-growing efficiency of not-so-intelligent machine intelligence – of specialised artificial intelligence, as opposed to general artificial intelligence.

This is a crucial point in the context of existential risk. We concede that AI poses significant risks, and we are not in denial of the potential for catastrophic risks (Rees, 2013). We are aware that it has been argued that the HBP could lead to machine consciousness or personhood (Lim, 2013).

However, we maintain that there is very little evidence that AI (at this point) poses an existential risk to the survival of humanity. However, it poses many manifest risks to particular parts of society, and that demands ways of understanding and mitigating such risks more effectively. It is some of the less speculative implications identified in the report referred to above (which may be no less potentially hazardous or disruptive) that this paper has therefore focused upon.

2.2 Near-term implications of AI

The important implications of AI that the Foresight Lab has flagged and examined, considering them to be of realistic near-term import, are broader issues and challenges. These are not specifically related to the research developed by the HBP, but to developments which several strands of research developed in the HBP have the potential to contribute to. They spring from developments in machine intelligence that we can see impacting already the following domains:

1. All areas involving affective relations between humans and intelligent machines, and in particular, care of vulnerable populations (the elderly, people with disabilities, children) and the sex industry.
2. The global labour market, with the replacement of the human workforce by intelligent machines, and subsequent economic impact.
3. All areas relying heavily on data analytics capabilities (e.g. marketing, insurance, credit scoring), with issues of data protection and privacy, data misuse and abuse, algorithmic opacity, etc.
4. Political, security, intelligence and military applications.
5. Energy consumption and electronic waste.

For a detailed overview of the points above, we refer the reader to (Rose et al., 2016), especially Section 3. 'Wider challenges and issues.' With regard to point 3 in particular, it is important to note that actions have already been taken by the HBP in acknowledgment of this area of concern. In response to an action plan formulated by SP12, the HBP has appointed a Data Protection Officer and supported the formation of a Data Governance Working Group which is in the final stages of creating a Data Policy Manual, in addition to the development of other resources for researchers and stakeholders by groups in SP12. Furthermore, work packages in SP12 are hosting a series of conferences on issues of Data Governance to bring together HBP members from across the project and external experts on related topics such as informed consent, privacy, data law, etc.

In view of these implications, the main high level conclusions drawn by the Foresight Lab for the HBP are summarized below.

2.3 Main findings and recommendations for action

1. In the short to medium term, a number of HBP Sub-Projects will (individually and collectively) contribute to the wider field of specialised machine intelligence. Many interdependencies between different parts of the project have become apparent, and through them, close relations between the fields of research and their respective practical domains. These deep interconnections should be attended to, especially as the cross-design projects started in the Operational Phase use these interconnections to construct research synergies. A systematic, project-wide reflection should be conducted to take stock of these synergetic potentials and devise a Responsible Research and Innovation roadmap for building on them.
2. Because of the strong potential for the HBP to contribute to the wider field of specialised machine intelligence in the short to medium term, there is a moral, if not legal, responsibility for HBP researchers and directors to work with the Ethics and Society Subproject (SP12) on how to address this challenge, and more generally to participate in current debates that address the need to make Artificial Intelligence ethical and socially beneficial.
3. There is a need for short term social and ethical issues of specialised machine intelligence to be demarcated from long term speculative risks associated to general machine intelligence. These short and medium term issues should be addressed as a priority, at national and trans-national levels, and the debates should not be left to the private sector and private philanthropy, but should be addressed openly and democratically. This is clearly a question playing out at a much broader level than the HBP, but one on which the HBP, as a Future and Emerging Technologies Flagship of the European Commission and contributor to developments in machine intelligence, could take a leading stance.
4. The human should not be 'designed out' of technology, instead the human should be put at the centre. A human-centred design approach that does not narrowly focus on the individual but takes into account the wider socioeconomic context, can bring to light a broader, and different, range of social and ethical issues. It is paramount that strategic choices and decisions driving research and innovation for future computing and robotics rely on such an approach.
5. No technology that uses the brain as its inspiration should neglect the capacity of the brain, in constant interaction with its interpersonal and physical environment, to develop and sustain a social and emotional mind. This is especially the case for applications in the domain of care (for older people, those with disabilities, children), which is a human interaction involving genuine reciprocation of feelings and obligations, and these entail the kinds of high level affective and interpersonal skills that are currently challenging for machine intelligence. Affective neuroscience and affective computing are areas of research of high relevance to machine

intelligence. They are at present minimally represented in the HBP, and could valuably complement the project.

6. In its Operational Phase starting in April 2016, the HBP moved into Horizon 2020, the EC funding programme expected to position Europe as a world-class competitor in ICT research and digital innovation through 'open science' and the development of innovative public-private partnerships. We recommend that those responsible for the scientific direction of the HBP set out policies that seek to ensure that the research results of public-private partnerships are subject to the same requirements of openness, to ensure that they are ethically sound and socially beneficial to European citizens. In particular, as the HBP proposes to offer a number of commercial services to industry through its infrastructure, an evaluation of proposed applications in terms of social and ethical impact should become an integral part of the terms of service.

2.4 Further conceptual issues

Beyond the high-level conclusions and recommendations formulated by the Foresight Lab, ethical and philosophical work in other parts of SP12 has identified some broad conceptual issues, which are briefly outlined below.

2.4.1 Human – machine interaction

The HBP pursues research in areas of neuro-mimetic computing and robotics as well as human cognition and its modelling. This latter dimension of HBP work includes reflection on how cognition can be investigated and understood, and reflection on the human through non-human means is not something to be treated lightly. As much as it is ethically important to critically assess research and development of artificial intelligence, or even agency, based on the assumptions manifested within it, it is equally important to challenge and scrutinise the fabric of those very assumptions (Collyer, 2011).

Technologies can embody (Bijker et al., 1987; Chander, 2016) and extend value agendas (Sandvig et al., 2015). This brings challenges in an acute sense where these technologies may be entrusted with areas of decision with human consequences – job automation, or AI-based market trading, for example – which can cause economic problems and social consequences.

Technology can have cognitive effects too, in augmenting or uprooting established assumptions and understandings about the world around us. Ambient intelligence, for instance, can challenge deep-seated assumptions about the objectivity of the environment (Schuurman et al., 2009), while more familiar technologies such as nuclear power, are argued to require particular political arrangements (Winner, 1980).

In turning the mirror on ourselves through the investigation of artificially intelligence and agency, there is a genuine opportunity to unearth, expose, and improve upon the tacit assumptions that underlie not just science and technology development, but also how we deal with the political, social, and cognitive dimensions of human beings.

2.4.2 'Human nature'

The Human Brain Project brings together methods from a range of disciplines with the intention of combining insights from all of them to understand or 'decode' the brain. From the molecule to the socialised organism, its account of the brain is aimed at bridging these scales. The benefits from this approach are hoped to come in areas like the diagnosis and treatment of brain diseases and psychiatric disorders.

Drawing upon other neuroscience and computing work such as the Blue Brain Project (Markram, 2006), a line can be seen emerging: from neuroscience data, neural activity is modelled in silico; from the in silico models, neuro-mimetic architectures are developed; from neuro-mimetic architectures, cognitive models are tailored for neuromorphic systems.

From this, where insight is sought in human cognition, it is important to remember the multiple abstractions and levels of modelling involved, which leaves a potential gap where context is discarded. Were the data collected and modelled to be considered as unvarnished representations of human cognition, it may appear to ground an objective *natural collectivity* (Couldry, 2014). This collectivity would be based in objective measurements such as neuronal spikings and synaptic activity. Such an approach would, at least tacitly, propose a 'we' that was reducible to these objective measurements. Moreover, in developing software architectures and systems based on the same data, it might be tempting to see their success as validation of such an objective collectivity, based in brain data. If data is collected from the brain, technologies are made based on the data, and the technologies seem to work well and comparably to a mind, it could easily seem that we have a valid reproduction of the very processes that gave rise to the data in the first place. It would be a tremendous risk to think in this way.

Any reduction will risk excluding some feature that might, even in another reduction, be important in explaining one phenomenon or another. It makes sense, then, to make reductions carefully, critically, and with fallibility in mind. We can gain great insights using data-centric approaches, and can develop fascinating technologies on the basis of such insights. But we ought to remember that these are not explanatory of human cognition across the board. The insights gained from abstract reductions are themselves abstract and are plausibly quite far removed from the actual cognitive lives of concrete human beings. Consequently, as we seek to apply insights from abstract reductions,

we must carefully reconstruct them in order to maximise what we can and cannot extrapolate from them, as Fuchs cautions in (Choudhury and Slaby, 2012, p. 331).

For instance, in the 'we' emergent from massive data, it is necessary to represent relevant features of cognition by proxy. The cognitive significance of neural goings-on might need to be inferred from sources other than those directly under scrutiny. Whereas sociology can resort to interview, fieldwork, and so on, to contextualise what is to be represented, a parallel resort is not clearly forthcoming in neuroscience. A risk here is that levels of explanation might be 'jumped' across in quite an opaque manner (Choudhury and Slaby, 2012, p. 311). Where simulation of the brain is cited as a method as well as an output from research (Amunts et al., 2013; Markram, 2006), the approach might seem to be justifying itself. Where neuroscience relies on computational and modelling practices as sources and methods, there is a risk that the brain becomes a kind of virtual entity that is presumed to hold answers to any question we might reasonably ask of it. It is not intuitively clear how to re-contextualise, and thereby to understand, what the significance of the outcomes of such an approach might be, and so we risk ad hoc justification of whatever we hope to establish. It is necessary therefore to reflect on the conceptual and epistemic challenges that attend this kind of large-scale, multi-disciplinary, and highly promising research endeavour.

While technology can tacitly embody values in an unquestioned way, and this can represent risks, human values can be eroded too in a context of rampant research and technology development. This represents a broad risk to the human self-image, in the sense that the constraining filter of technology risks reducing the richness of the human to an impoverished and distorted version of itself.

3 Closing the anticipation-action loop: The role of ethics management and researcher awareness

We will now present how the high level conclusions and recommendations formulated by the Foresight Lab, along with other broad conceptual issues identified by ethical and philosophical work in other parts of SP12, are being taken up by the researcher awareness and ethics management teams, to be developed into a concrete action plan for the HBP.

The Ethics and Society section of the HBP (SP12) contains multiple interrelated processes that can facilitate and foster reflexivity within a project, which can help relevant stakeholders to build capacity against the potential issues, challenges, and other impacts of technology development, such as AI.

The overall aim of Ethics Management in tandem with Researcher Awareness and as part of the RRI activities in the HBP, is to develop a research ecosystem that empowers all stakeholders to reflect on the work undertaken in the project, their role in it, its justification, and its potential implications. In the following section we describe the structures put in place to achieve this and their application in practice.[9]

3.1 Ethics management and researcher awareness

Ethics Management has set up an organisational structure that aims to foster collaboration and reflection. Key components of these are the Ethics Advisory Board, the Ethics Rapporteur Programme, the Point Of Registration for Ethics concerns (PORE), and Ethics Compliance.

The Ethics Advisory Board (EAB)[10] is comprised of independent experts in the various ethical issues of the HBP, such as animal research, human subject research, biobanks, or data protection. It also includes experts on robotics and human-technology interaction with significant experience in the ethics of AI. Members of the EAB are appointed based upon their competency in a relevant subject in order to broadly cover the areas researched within the HBP, with new appointments ratified by the Stakeholder Board of the HBP. This advisory board, a common project governance structure, is complemented by the Ethics Rapporteur Programme[11]. Ethics Rapporteurs are researchers who are nominated by each of the 11 scientific and technical subprojects[12] of the HBP to specifically deal with and disseminate information regarding ethical issues. By establishing this programme, it was possible to collect the required subject expertise and ensure that each technical community has a point of contact with regard to ethical issues. Ethics Advisory Board members and Ethics Rapporteurs are paired up to ensure a reciprocal flow of information. Both groups work closely with Ethics Management to update the current state of ethical issues.

A final relevant project structure is the Point Of Registration for Ethics concerns (PORE). This is a mechanism by which anybody, both within or outside the HBP, can register an ethical concern. It is implemented as a simple web-based survey that allows people to contact the Ethics Management team, either using their names or anonymously. Issues registered with PORE are then investigated and addressed according to their nature through the compliance process or other aspects of the

[9] See the Ethics Management webpage for more infohttps://www.humanbrainproject.eu/en/social-ethical-reflective/ethics-resources/, accessed 24/11/2017

[10]https://www.humanbrainproject.eu/en/open-ethical-engaged/ethics/ethics-advisory-board/ , accessed 24/11/2017

[11]https://www.humanbrainproject.eu/en/open-ethical-engaged/ethics/ethics-rapporteurs/ , accessed 24/11/2017.

[12] This includes subproject SP12-Ethics and Society, which develops research activities alongside its ethics management and public engagement activities.

ethics governance framework, potentially drawing upon the knowledge of ethical issue type specialists in the Ethics Management team (e.g. personal data, animal data, human data).

Ethics Management also has a strong compliance management component which works with the individual researchers and principal investigators in the project to ensure that they have the appropriate approvals for their work. Ethics compliance collects such documentation and makes it available to the European Commission and its ethics reviewers.

Researcher Awareness, the final organisational structure to be introduced here¸ aims to reach out to the membership of the HBP and broaden the scope of discourse and reflection. It is tasked with developing mechanisms that allow the insights gained by the Ethics and Society subproject of the HBP (SP12) to be turned into practice in the scientific and technical sections.

Organisational Structure Components	**Governance Aspect or Relationship**	**Description**
Ethics Advisory Board (EAB)	Advise HBP Directorate and Science and Infrastructure Board	Independent body
Ethics Rapporteur Programme	Communicate with EAB and Ethics Management Disseminate ethics information to their subproject	Members of each scientific or technical HBP subproject
Point Of Registration for Ethics concerns	Accept submission of ethics concerns from the public or project members	Online public portal
Ethics Compliance	Collect ethical approval documentation Communicate with EC and reviewers	Subject area experts in Ethics Management
Researcher Awareness	Develop Ethics and Society insights into practice Communicate with membership of the HBP	Members of the Ethics and Society subproject

Ethics Management and Research Awareness: Organisational structure components

Researcher Awareness and Ethics Management utilise a set of tools including surveys, interviews, meetings, and workshops in order to create a discursive environment in which ethical, social, and other issues can be identified and debated. These can be related to regulatory frameworks, disciplinary norms, anticipated near or far future potentialities, or other such framing ideas. Once

identified, issues that can be addressed by compliance processes are so dealt with, according to EC standards. Aside from this, SP12 research activities are brought to bear more acutely upon HBP-wide research. This permits attention to be paid to novel or unforeseen issues which can then be productively engaged with and eventually incorporated into regulatory frameworks.

The intention of this system is to go beyond mere ethico-legal compliance and the unquestioned assumptions of received disciplinary wisdom to open up a space for thoughtful reflection during the course of ongoing research. This reflection is intended to be transformative.

3.2 'Open loop'

There is an inherent challenge in making transformative interventions in research. The 'anticipate, reflect, engage, and act' (AREA) framework serves well as a way to frame this transformative challenge. The issue just described comes in the final 'A' of the framework. Whilst it is clear that anticipation, reflection, and engagement take place, it is less obvious how this manifests in action. Imaginative strategies are needed that will 'close the loop' (although a more appropriate metaphor might be that of a spiral unfolding in time). The way the open loop challenge is currently addressed in the HBP includes at least four complementary components:

a) action plans
b) workshops
c) leveraging the ethics rapporteur programme
d) using existing ethics management tools

Researcher awareness work synthesises action plans, based on original research and on the insights gained into the workings of all parts of the HBP. These action plans can be developed and tested through various kinds of activities, like webinar and workshop environments, as well as more experimental formats. They can both inspire the topics of activities and be refined in those same activities. Workshops which have been tailored for the target audiences (e.g. members of a task or laboratory; researchers focusing on a common topic, or at specific career stages) and framed around specific issues have proven successful as a means of engagement between Ethics and Society and the rest of the HBP. As an example of such a precisely tailored workshop, in November 2017 the Foresight Lab co-organised a workshop with the Neuromorphic Computing subproject (SP9), focusing on practical ethical questions arising within this particular domain of research (ethical concerns linked to dual-use, in the context of an open science policy and of collaboration with industry to promote innovation).

But 'closing the loop' through action should concern just as much the research practices of the social scientists as those of the various other scientists and engineers working in the Human Brain Project. Taking on board the outcomes of past and ongoing experience, the Foresight Lab is thus conducting methodological research and experimenting with different formats of engagement. For instance, collaborating with scientists, engineers, and science fiction writers to turn current lab research into near-future science fiction stories. An example of this was working with Alan Winfield,[13] Professor of Robot Ethics at the University of West England Bristol and member of the Ethics Advisory Board of the HBP, the Bristol Robotics Laboratory (BRL)[14] and writers Stephen Oram,[15] Allen Ashley[16] and Jule Owen.[17] During a public event hosted by publisher SilverWood Books[18] at the Bristol Festival of Literature in October 2016, each read a 'near future' fictional short story that they had written following their visit and exchanges with BRL researchers. The stories were then used as springboard for a moderated discussion between the authors, a panel of BRL scientists, and the audience.[19] Further initiatives are in development, notably aimed at building awareness and self-reflexive capacity in early career scientists, around their research and their role in society.

Once clear action plans are devised in response to issues of concern, they are disseminated across the HBP among (at a minimum) the existing ethics rapporteurs, who are typically researchers with an interest in multidisciplinary and interdisciplinary approaches to research. In providing this group with appropriately contextualised action plans based in solid and tested research, researcher awareness demonstrates the outcomes of research, and at the same time increases the capacity of the rapporteur group – providing them with the means to frame and re-frame existing approaches to research. This is highly valuable, especially within the general context of responsible innovation throughout European research.

To approach more nebulous potential areas of ethical and social impact, such as the role of advanced AI in a possible future beyond the HBP, it is desirable that researchers from across different disciplines, as well as experts from outside academia, are brought together in appropriate fora. Each of these must be based in a well-defined area of interest and use methods that can address them. In terms of topics, these are identified through targeted SP12 research such as that contained in Foresight Reports, through ethics self-assessments made by individual subprojects and

[13] http://people.uwe.ac.uk/Pages/person.aspx?accountname=campus%5Ca-winfield, consulted 20/12/2017.
[14] http://www.brl.ac.uk/, consulted 20/12/2017.
[15] http://stephenoram.net/, consulted 20/12/2017.
[16] http://www.allenashley.com/, consulted 20/12/2017.
[17] https://juleowen.com/, consulted 20/12/2017.
[18] http://www.silverwoodbooks.co.uk/, consulted 20/12/2017.
[19] A full recording of the event can be watched at https://www.youtube.com/watch?v=8HOLCH7H1rs.

coordinated by the Ethics Rapporteurs, or through Opinions derived from SP12 collaborative work. As discussed, this is where the focus on intelligent machines, for instance, is derived.

Addressing these issues typically involves providing or otherwise commissioning training events for researchers working in relevant areas, such as high performance computing, neuromorphic computing or neurorobotics. In order to ensure accurate, well-delivered, and up-to-date training, consultancy may be used in order to get high-quality external expertise involved in the delivery of workshops and so on. This should become normalised in the HBP such that it becomes standard practice.

Ethics rapporteurs have expressed a strong desire for concise, targeted, relevant, and manageable actions. For this reason, workshop formats, 'brown-bag lunches', webinar series, and other such pedagogical devices are preferred to large conference-type approaches. This prevents over-saturation and therefore maximises impact, while maintaining a discursive ideal. This is also in line with the general principles of Ethics and Society approaches (Stahl et al., 2016), which must be considered as part of an HBP-wide dialogue over time. This also relieves the pressures associated with thinking of the actions as 'one stop shops' that might fuel a tendency to overdo content. Finally, this expression of preference over the style of communication should also prompt recognition of the need for skilled, qualified, and experienced action providers.

3.3 Challenges to closing the loop

Many challenges exist with regard to closing the AREA loop in the context of the HBP. Amongst these are the complexities of effective communication and action plan implementation across such a large project, and the conceptual framing of SP12, Ethics and Society within the HBP (see Aicardi et al., 2017). Furthermore, across a project as topically complex and as international as the Human Brain Project, it is to be expected that different perspectives will arise and harbour potentially divergent interests and values. We need not go to the level of the individual researcher to make this point, although it is relevant, but the level of disciplinary differences is sufficient for the purpose of illustration. Highly trained and skilled researchers are employed for diverse kinds of research across all the various dimensions of the HBP. Each, in being expert in their field, brings with them a pedigree of situational understanding grounded within their own domain of activity, or their 'disciplinary matrix' (Kuhn, 1996).

Disciplinary matrices are predicated on long historical developments of aims, practices, norms, and standards. They may have much in common with one another, formally or generatively, yet they diverge in concrete and conceptual terms. Bringing together those operating in different arenas thus represents a challenge, not least in terms of the legitimacy of any intervention. Interdisciplinary,

interventionist governance predicated on ethics, or on other transformative-by-intention grounds, could be said to have problems of perceived legitimacy, in the vein of those envisaged by socio-political analysts (Habermas, 1980, 1974).

These tendencies can emerge owing to the nature of intervention in disciplinary matrices, which essentially involves the questioning of factors that are often tacitly accepted as norms. Opening 'taken-for-granteds' (Hopper, 1981) to discursive scrutiny can appear as an affront, or a potentially arbitrary displacement of one value for another, with associated perceived threats to domains of expertise and disciplinary integrity.

At a very abstract level, the possibility of transforming ongoing interdisciplinary research according to RRI principles is predicated on this sort of discursive widening of norms and values among diverse research groups. Despite the difficulties this can pose, it can nevertheless be of tremendous value as it can produce an overall research context within which the researchers themselves organise their own steering parameters, and do so beyond constraints received from narrower traditional approaches, or defined by regulatory frameworks. In effect, the researchers themselves, allied and in conjunction with ethics and researcher awareness colleagues, ask and address a question analogous to the following:

"How would the members of a social system, at a given stage in the development of productive forces, have collectively and bindingly interpreted their needs (and which norms would they have accepted as justified) if they could and would have decided on the organization of social intercourse through discursive will formation, with adequate knowledge of the limiting conditions and functional imperatives of society?" (Habermas, 1980, p. 113)

Rather than discussing a social system writ large, instead we have this challenge in a social system constrained by a given global aim, the HBP mission, and comprised of multiple sets of disciplinary matrices. So the question becomes:

How would a multi- and inter-disciplinary group of researchers have collectively and bindingly interpreted their needs and norms for action if they could have decided on their global agenda for interaction, given their knowledge of overall aims and external expectations?

This is a very big goal, and one which is complex to pursue. Our efforts to recapture and reformulate the discursive consensus of a research community, against the imperatives of a globalised research-industrial syndrome, may just provide explanatory strategies that can be deployed to further undermine an already depoliticised public conceived as an obstacle to transcendent technological progress. Scientific discourse in general can all too easily gain control through instrumental

successes, but at the cost of meaning. This model can tend to leave the rightness, appropriateness, or the value in general, of the research task at hand appear unproblematic, but it is only so in the unquestioned value context of a future already assumed to be worth living (Habermas, 1996) or already set in stone.

Through the use of ethics and society research, and the dialogical governance methods described above, the HBP has sought to internalise the sorts of response necessitated by these general risks. With this approach, the intention is that specific x-risks such as those potentially posed by AI, are mitigated in the project lifecycle itself, and also more broadly in building reflective capacity in the researchers. This is hoped to facilitate a culture change in research fields such that reflection on x-risks, and matters in general beyond mere compliance, are routinely considered in the course of ongoing scientific research.

3.4 Mitigation

The challenge presented by many of the conclusions, recommendations and other conceptual issues that we have presented before in part 2 is that they represent broad statements of often abstract dimensions of reflection upon HBP research that are not readily actionable in the day to day activities of a lab or research group. Nevertheless, they represent loci of thought that can have real impact in the future. Whether or not in terms of the specifics envisioned, they encapsulate animating features of futures that could be enabled by present research. More importantly than predictive accuracy, however, they represent an opportunity to bring curiosity into research. Doing this successfully can serve to reorient research practice from a position potentially divorced from the context of a value laden social world, into an activity responsible to that social world.

In these contexts, the AREA framework and the activities of Ethics and Society research and governance in the HBP can be seen to make a crucial difference. The first challenge is to differentiate between ethics compliance and broader reflection upon ethics and society. This is achieved in the HBP through a demarcation between compliance activities, which are structured according to a self-assessment survey and a round of approval-checking meetings, and subsequent activities with a broader mandate. Teleconferences and other forms of interpersonal interaction involving researchers, ethics rapporteurs, EAB members, Ethics Management, and Researcher Awareness representatives serve to open the framing from ethics compliance to this broader reflection.

For example, in terms of ethics compliance, a Terminator-style robot is not a plausible concern for the outcomes of the HBP. But in a broader reflection upon how AI and robotics intersect, researchers are keen to consider how their foundational work could appear in a less immediate yet not so distant future scenario: what responsibilities would they bear for their work, even if it were

mis/re-appropriated? How far into the future ought they be expected to anticipate? A formalised compliance context may not permit this reflection, and so it is incumbent that the space is opened by other means.

Meanwhile, broadly reflecting on neuro-mimetic computing and modelling techniques can ground an approach to the investigation that does not become mired in a too-naive faith in a single interpretation of data. This can serve to encourage the investigation of a variety of areas where the focus is on broader interpretation (Boyd and Crawford, 2012). This can only benefit the quality of the science produced. As long as a space is opened for detailed reflection upon every dimension of research from idea, to lab bench, to possible implications, these benefits are available.

The point of opening this space is not to answer these reflections once and for all – it would be faintly ridiculous to expect a date for 'the singularity', or when models will complete neuroscience – but instead to normalise the deepening of critical reflection prior to, and in the course of, ongoing research. The building of this capacity in research programmes, making the space for it, increases the capacities of the researcher and improves their research. Ideally, even when not faced explicitly with a public audience, research is always being interrogated by norms and values not intrinsic to the research underway. This gives the research a self-critical dimension that would otherwise be missing (Lavelle and Rainey, 2013). In terms of x-risks, this constructs the issues as research flows, thereby bringing related ideas within the horizon of awareness for consideration rather than having them emerge as a shock.

4 Conclusion: Benefits and limitations of the methodological approach

Our practices highlight the possibilities of implementing reflexive processes within large, interdisciplinary projects and how these can bring serious research risks to attention across a variety of dimensions. It is also possible to implement processes to treat such issues in a way that recognises multiple timescales, and (sensibly constrained) speculation about possible futures. These processes, being centred upon researchers as actors, aim to transform not just research tasks, but the research community in order to generate a research culture in which reflection is a norm. This reflection is encouraged beyond the lab, with research conceptualised in a more global context.

Limitations of this approach include its complexity, resource intensiveness, and requirement of commitment from all involved. As it flies somewhat in the face of standard practices concerning, say, ethics (research design, ethics approval, research) in seeking to be transformative and ongoing, this commitment cannot be assumed. Trust, enthusiasm, and understanding therefore need to be built

early, and actively maintained throughout the lifetime of a project. This adds another element of time, which is often difficult to fit into research.

Another, extrinsic limitation comes in the form of research funding instruments in general. Any attempt at research governance, steering, or conditioning that seeks to be transformative will struggle in contexts that insist upon fine-grained specification of research programmes in advance of funding. Where milestones, deliverables, and other such outputs must be specified along with methods in advance, it can stifle the possibility of genuine reflection in that research. A possibility for minimising this problem, thus giving the opportunity for ethical reflection to play a meaningful role in a project's progression, might be a stage-gate system of evaluation (Cooper, 1990) such as that used for the Stratospheric Particle Injection for Climate Engineering (SPICE) project in the UK in the 2000s (Stilgoe et al., 2013), provided it is not enacted as a box-ticking exercise. Yet any procedure at the project level, even when fully embedded within the research project, will be constrained in its effectiveness because the research often depends completely upon the funding regime in which it appears. This includes whatever tacit, unquestioned norms accompany the relevant funding framework.

Following from the previous point, there are indeed limitations on undertaking responsible research and innovation exclusively at the project level. By definition, the individuals involved are researchers in the various fields who are (in most cases) positively inclined towards these fields. They understand the subject matter and are therefore invaluable as sources of information about the technology in question. At the same time, they may have a positive bias toward the technology and view any perceived risks as less dramatic than outsiders might. From the perspective of existential risks, this means that the work on the project level that we have described here must be supplemented with work on a higher level that can not only take a more detached view of the research in question, but also situate it against a wider, changing socio-political context. This tallies, in fact, with some of the higher level conclusions and recommendations of the Foresight Lab presented in section 2.3.

Another limitation at the project level is what we have described elsewhere as the ‘synchrony mirage,’ inherent to visions of upstream embedding of RRI into research projects. Research projects are hardly self-contained, as research teams usually develop and pursue a research agenda across a succession of overlapping projects. This is indeed the case within the HBP, where a number of its core research strands began well over a decade ago. As a result, research and innovation may have advanced substantially along previously-set trajectories from the start of a particular project. This strongly constrains the deployment and effectiveness of RRI practices (Aicardi et al., 2017, p. 14).

The challenge is, how is it possible to overcome these limitations and allow ethical considerations to bear more actively on research directions? Our combined experience suggests that there is a case for RRI to be implemented at multiple levels beyond that of the individual project, like that of a research funding programme or even as a cross-funding activity. Yet acting on these levels still does not address the content of research & innovation frameworks themselves. We would suggest that there is a case for deploying RRI practices at the level of the elaboration and management of research & innovation policies and funding programmes, thus also opening them up to the possibility of change through ongoing ethical and critical reflection.

Having said this, the paper has highlighted that the HBP has the potential to be an important player in the further development of AI, and that consideration of possible risks should be undertaken. It has shown how, accordingly, the HBP Ethics and Society team has been deploying a RRI framework to accompany technology development in the project. The structures of Responsible Research and Innovation that we have described show how the theoretical aims of RRI can be applied in practice. We have shown that it is possible to create and use project governance structures that are conducive to risk assessment and management. This, we believe, is the major, highly practical, contribution of this paper.

5 References

Aicardi, C., Reinsborough, M., Rose, N., 2017. The integrated Ethics and Society Program of the Human Brain Project: Reflecting on an ongoing experience. J. Responsible Innov.

Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.-É., Bludau, S., Bazin, P.-L., Lewis, L.B., Oros-Peusquens, A.-M., Shah, N.J., Lippert, T., Zilles, K., Evans, A.C., 2013. BigBrain: An Ultrahigh-Resolution 3D Human Brain Model. Science 340, 1472–1475. https://doi.org/10.1126/science.1235381

Bijker, W., E., Hughes, T.P., Pinch, T. (Eds.), 1987. The social construction of technological systems: new directions in the sociology and history of technology. MIT Press, Cambridge, Mass.

Boyd, D., Crawford, K., 2012. Critical Questions For Big Data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15, 662–679. https://doi.org/10.1080/1369118X.2012.678878

Chander, A., 2016. The Racist algorithm?

Choudhury, S., Slaby, J., 2012. Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience. Wiley-Blackwell, Chichester, UK.

Collyer, F., 2011. Reflexivity and the Sociology of Science and Technology: The Invention of“ Eryc” the Antibiotic. Qual. Rep. 16, 316.

Committee on Legal Affairs, 2017. REPORT with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)) (No. A8–0005/2017). European Parliament.

Cooper, R.G., 1990. Stage-gate systems: A new tool for managing new products. Bus. Horiz. 33, 44–54. https://doi.org/10.1016/0007-6813(90)90040-I

Couldry, N., 2014. Inaugural: A necessary disenchantment: myth, agency and injustice in a digital world. Sociol. Rev. 62, 880–897. https://doi.org/10.1111/1467-954X.12158
Evers, K., 2017. The contribution of neuroethics to international brain research initiatives. Nat Rev Neurosci 18, 1–2. https://doi.org/10.1038/nrn.2016.143
Executive Office of the President, 2016a. Preparing for the Future of Artificial Intelligence. Executive Office of the President National Science and Technology Council Committee on Technology.
Executive Office of the President, 2016b. Artificial Intelligence, Automation, and the Economy. Executive Office of the President National Science and Technology Council Committee on Technology.
Habermas, J., 1996. Between facts and norms: contributions to a discourse theory of law and democracy, Studies in contemporary German social thought. MIT Press, Cambridge, Mass.
Habermas, J., 1980. Legitimation crisis. Heinemann, London.
Habermas, J., 1974. Theory and Practice. Beacon Press, Boston.
Hopper, R., 1981. The Taken-for-Granted. Hum. Commun. Res. 7, 195–211. https://doi.org/10.1111/j.1468-2958.1981.tb00569.x
House of Commons Science and Technology Committee, 2016. Robotics and artificial intelligence.
Kuhn, T.S., 1996. The structure of scientific revolutions, 3rd ed. ed. University of Chicago Press, Chicago, IL.
Lavelle, S., Rainey, S., 2013. Transformation of Proceduralism from Contextual to Comprehensive. Httpservicesigi-Glob.-1-4666-3670-5ch021 312–343. https://doi.org/10.4018/978-1-4666-3670-5.ch021
Lim, D., 2013. Brain simulation and personhood: a concern with the Human Brain Project. Ethics Inf. Technol. 1–13. https://doi.org/10.1007/s10676-013-9330-5
Markram, H., 2006. The Blue Brain Project. Nat. Rev. Neurosci. 7, 153–160. https://doi.org/10.1038/nrn1848
Rainey, S., Stahl, B., Shaw, M.C., Reinsborough, M., 2017. Ethics Management and Responsible Research and Innovation in the Human Brain Project, in: von Schomberg, R. (Ed.), Handbook - Responsible Innovation: A Global Resource. Edward Elgar Publishing Ltd., Cheltenham.
Rees, M., 2013. Denial of Catastrophic Risks. Science 339, 1123–1123. https://doi.org/10.1126/science.1236756
Rose, N., Aicardi, C., Reinsborough, M., 2016. Foresight Report on Future Computing and Robotics. An Ethics and Society deliverable of the Human Brain Project to the European Commission.
Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C., 2015. Can an Algorithm be Unethical? Ann Arbor 1001, 1285.
Schuurman, J.G., El-Hadidy, F.M., Krom, A., Walhout, B., 2009. Ambient Intelligence–Viable future or dangerous illusion. Hague Rathenau Inst.
Stahl, B.C., Rainey, S., Shaw, M., 2016. Managing Ethics in the HBP: A Reflective and Dialogical Approach. AJOB Neurosci. 7, 20–24.
Stilgoe, J., Owen, R., Macnaghten, P., 2013. Developing a framework for responsible innovation. Res. Policy 42, 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008
Weingart, P., 2010. Science, the Public and the Media – Views from Everywhere, in: Carrier, M., Nordmann, A. (Eds.), Science in the Context of Application. Springer, New York.
Winner, L., 1980. Do artifacts have politics? Daedalus 121–136.